

*To our beloved Sensei
Prof. Susumu Horiguchi*

Acknowledgments

All praises goes to Almighty Allah (swt), the most merciful and sustainer, who has been showering his endless blessings on the author throughout of his life.

The authors are very much pleased to express my sincere and profound gratitude and profuse heartfelt thanks to late Prof. Susumu Horiguchi of GSIS at Tohoku university for his constant encouragement and kind guidance during this research work. We are deeply grateful to him for his enthusiasm and insight, which have made research a most enjoyable and fulfilling experience. His phenomenal depth of knowledge and ability to discern the key points of the research problem inspire us a lot. We are both incredibly proud and immensely lucky to become his students.

We have learned much from discussion with colleagues over the years including Prof. Y. Hibino (JAIST), Prof. M. Kaneko (JAIST), Dr. Y. Miura (Shonan IT), Dr. M. Fukushi (GSIS, Tohoku Univ), Dr. Y. Fukushima (Sophia Univ., Tokyo), Dr. R. Hayashi (Kanazawa IT), E. Horiguchi (JAIST) for their helpful discussion, comments, and suggestions during this research work. The authors are indebted to the to the Ministry of Education, Science, Sports, and Culture, Japan for the financial support of this research work.

Last but not least, the authors are heartfelt thankful to their family members for their bountiful forbearance and significant sacrifice and encouragement. They were putting up with many dreary evenings and weekends when we were doing this research work. Their tolerance and understanding made it possible for us to complete this endeavor.

Contents

Acknowledgments	ii
1 Introduction	1
1.1 Introduction	1
1.2 Interconnection Networks	2
1.3 Motivations and Goal	6
1.4 Contribution of the Book	8
1.5 Synopsis of this Book	9
2 Interconnection Networks for Massively Parallel Computers	11
2.1 Introduction	11
2.2 Definitions	13
2.2.1 Fundamental Definitions	13
2.2.2 Topological Characteristics of Interconnection Networks	13
2.2.3 Layout Characteristics of Interconnection Networks	14
2.2.4 Dynamic Communication Performance Metrics	14
2.3 Interconnection Network Topologies	14
2.3.1 Completely-Connected Networks	15
2.3.2 Star Networks	15
2.3.3 Tree Networks	16
2.3.4 Hypercubic Networks	19
2.3.5 Array Networks	21
2.4 Hierarchical Interconnection Network (HIN)	24
2.4.1 Completely-Connected Network based HIN	25
2.4.2 Tree Network based HIN	27
2.4.3 Hypercube Network based HIN	28
2.4.4 Array Network based HIN	30
2.5 Conclusions	35
3 Hierarchical Torus Network (HTN)	36
3.1 Introduction	36
3.2 Architecture of the HTN	37
3.2.1 Basic Module	37
3.2.2 Higher Level Interconnection	38
3.2.3 Addressing and Routing	40
3.3 Static Network Performance	43

3.3.1	Node Degree	43
3.3.2	Diameter	44
3.3.3	Average Distance	46
3.3.4	Cost	48
3.3.5	Connectivity	48
3.3.6	Bisection Width	49
3.4	Wafer Stacked Implementation	51
3.4.1	3D Stacked Implementation	51
3.4.2	Peak Number of Vertical Links	51
3.4.3	Layout Area	56
3.4.4	Maximum Wire Length	60
3.5	Conclusions	62
4	Dynamic Communication Performance of the HTN	63
4.1	Introduction	63
4.2	Routing Algorithm	64
4.2.1	Resources and Allocation Units	65
4.2.2	Taxonomy of Routing Algorithm	66
4.2.3	Primitive Considerations	67
4.2.4	Channel Dependency Graph	73
4.3	Dimension-Order Routing (DOR) for HTN	76
4.3.1	Routing Algorithm for HTN	77
4.3.2	Deadlock-free Routing	79
4.3.3	Minimum Number of Virtual Channels	82
4.4	Dynamic Communication Performance using DOR	83
4.4.1	Performance of Interconnection Networks	83
4.4.2	Simulation Environment	84
4.4.3	Traffic Patterns	85
4.4.4	Dynamic Communication Performance Evaluation	88
4.4.5	Effect of Message Length	102
4.4.6	Effect of the Number of Virtual Channels	103
4.5	Adaptive Routing	105
4.5.1	Link-Selection (LS) Algorithm	106
4.5.2	Channel-Selection (CS) Algorithm	108
4.5.3	Combination of LS and CS (LS+CS) Algorithm	110
4.5.4	Deadlock-Free Routing	110
4.6	Router Cost and Speed	113
4.6.1	Router Gate Counts	113
4.6.2	Router Speed	114
4.7	Dynamic Communication Performance using Adaptive Routing	117
4.8	Conclusions	122
5	Reconfiguration Architecture and Application Mappings	125
5.1	Introduction	125
5.2	Reconfiguration Architecture of the HTN	126
5.2.1	Reconfiguration Scheme	126

5.2.2	System Yield of the HTN	127
5.3	Application Mappings on HTN	131
5.3.1	Converge and Diverge	132
5.3.2	Bitonic Merge	132
5.3.3	Fast Fourier Transform (FFT)	134
5.3.4	Finding the Maximum	135
5.3.5	Processing Time	135
5.4	Conclusions	141
6	Pruned Hierarchical Torus Network	142
6.1	Introduction	142
6.2	Pruned Network	143
6.2.1	Pruned Torus Network	143
6.2.2	Pruned HTN	145
6.3	3D-WSI Implementation of the Pruned HTN	147
6.3.1	Peak Number of Vertical Links	147
6.3.2	Layout Area	147
6.4	Conclusion	149
7	Modification of other Hierarchical Networks based on Torus-Torus Interconnection	150
7.1	Introduction	150
7.2	Modified Hierarchical 3D-Torus Network	150
7.2.1	Interconnection of the MH3DT Network	151
7.2.2	Routing Algorithm	153
7.2.3	Deadlock-Free Routing	155
7.2.4	Static Network Performance	156
7.2.5	Dynamic Communication Performance	158
7.2.6	Summary	162
7.3	Modified TESH Network	163
7.3.1	Interconnection of the TTN	163
7.3.2	Routing Algorithm	165
7.3.3	Deadlock-Free Routing	167
7.3.4	Static Network Performance	168
7.3.5	Dynamic Communication Performance	169
7.3.6	Summary	172
7.4	Conclusion	172
8	Conclusions	173
8.1	Introduction	173
8.2	Conclusions	173
8.3	Future Directions	176
	References	178
	Index	191

List of Figures

2.1	Completely-connected networks for $N = 4$, $N = 8$, and $N = 12$.	15
2.2	(a) A star-connected network of nine nodes (b) Star graph network	16
2.3	A 15 node binary tree network	17
2.4	A 15 node X-tree network	17
2.5	The 2D mesh-of-trees. Leaf nodes from the original grid are denoted with black circles. Nodes added to form row trees are denoted with red squares, and nodes added to form column trees are denoted with blue squares	18
2.6	A fat-tree network	18
2.7	The Binary cube networks of zero, one, two, three, and four dimensions, the nodes are labeled using n -bit binary numbers.	20
2.8	(a) The 3-dimensional binary cube network (b) The 3-dimensional CCC. Labels for individual nodes in the CCC are binary cube node label and the adjacent link label.	21
2.9	A four-node linear array and ring network	22
2.10	A layout for a ring network which minimizes link lengths ($N = 8$).	22
2.11	2D mesh and torus networks with 4 nodes in each dimension	23
2.12	3D mesh and torus networks with 4 nodes in each dimension	24
2.13	A level-2 MFC network with 8 clusters and the cluster size is 8.	26
2.14	An example of 16-node swapped network with the 4-node complete graph as its basis.	26
2.15	A pyramid network of 16 node	27
2.16	A hierarchical clique network	28
2.17	Fibonacci cubes	29
2.18	A HCN(2,2) network	30
2.19	Recursive diagonal torus network	31
2.20	Standard 1D-SRT consisting of 32 nodes.	32
2.21	Level-2 interconnection of TESH network	33
2.22	Interconnection of a Level-2 H3D-torus network	34
2.23	Interconnection of a Level-2 H3D-mesh network	35
3.1	Interconnection of HTN	37
3.2	Basic module of the HTN	38
3.3	Interconnection of a Level-2 HTN	39
3.4	Interconnection of a Level-3 HTN	39
3.5	Routing algorithm of the HTN	42
3.6	Illustration of degree of HTN	43
3.7	Diameter of networks as a function of number of nodes (N)	45
3.8	Average distance of networks as a function of number of nodes (N)	47

3.9	Average distance of various networks with 4096 nodes.	47
3.10	Cost of different networks as a function of number of nodes (N)	48
3.11	Illustration of connectivity for 2D-mesh network.	49
3.12	Bisection width of networks as a function of number of nodes (N)	50
3.13	Structure of 3D stacked implementation	52
3.14	Structure of microbridge and feedthrough	52
3.15	PE array in a silicon plane for wafer stacked-implementation	53
3.16	Vertical links of 2D-mesh network in 3D wafer stacked-implementation	53
3.17	Vertical links of 2D-torus network in 3D wafer stacked-implementation	53
3.18	Interconnection scheme of 2D-torus in 3D stacked implementation	54
3.19	A comparison of peak number of vertical links of HTN with other networks	56
3.20	Layout area of 2D-torus for $N = 16$, $L = 4$ and $p = 1$	58
3.21	Normalized layout area	60
3.22	2D-planner realization of 3D-torus network.	61
4.1	Units of resource allocation.	66
4.2	Wormhole routing	68
4.3	An example of the blocked wormhole-routed message	68
4.4	Time-space diagram of a wormhole-routed message	69
4.5	An example of deadlock involving four packets	70
4.6	Virtual channel	71
4.7	Message blocking while physical channels remain idle	72
4.8	Virtual channel allows to pass blocked message	72
4.9	(a) A ring network with unidirectional channels. (b) The associated channel dependency graph contains a cycle. (c) Each physical channel is logically split into two virtual channels. (d) A modified channel dependency graph without cycles.	74
4.10	Deadlock configuration in (a) mesh network (b) torus network	76
4.11	A set of routing paths created by the dimension order routing in a 2D-mesh network	77
4.12	Dimension-order routing algorithm for HTN	80
4.13	An example of message routing in HTN	81
4.14	Nonuniform traffic patterns on a 8×8 mesh networks: (a) dimension-reversal traffic and (b) bit-reversal traffic	87
4.15	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: (a) 1024 nodes, different virtual channels, short message, and $q = 0$ (b) 1024 nodes, 3 virtual channels, short message, and $q = 0$	90
4.16	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, short message, and $q = 1$ (b) 512 nodes, 3 virtual channels, short message, and $q = 1$, (c) 1024 nodes, 3 virtual channels, short message, and $q = 1$, (d) 1024 nodes, 3 virtual channels, medium-length message, and $q = 1$, (e) 1024 nodes, 3 virtual channels, long message, and $q = 1$,	91

4.17	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: (a) 256 nodes, 2 virtual channels, short message, and $q = 1$ (b) 256 nodes, 2 virtual channels, medium-length message, and $q = 1$, (c) 256 nodes, 2 virtual channels, long message, and $q = 1$, (d) 1024 nodes, 2 virtual channels, short message, and $q = 1$, (e) 1024 nodes, 2 virtual channels, medium-length message, and $q = 1$, and (f) 1024 nodes, 2 virtual channels, long message, and $q = 1$	92
4.18	Dynamic communication performance of dimension-order routing with hot-spot traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, 5% hot-spot traffic, short message, and $q = 1$ (b) 512 nodes, 3 virtual channels, 5% hot-spot traffic, short message, and $q = 1$, and (c) 1024 nodes, 3 virtual channels, 5% hot-spot traffic, short message, and $q = 1$. . .	94
4.19	Dynamic communication performance of dimension-order routing with hot-spot traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, 2% hot-spot traffic, short message, and $q = 1$ (b) 256 nodes, 3 virtual channels, 10% hot-spot traffic, short message, and $q = 1$, (c) 1024 nodes, 3 virtual channels, 2% hot-spot traffic, short message, and $q = 1$, and (d) 1024 nodes, 3 virtual channels, 10% hot-spot traffic, short message, and $q = 1$	95
4.20	Dynamic communication performance of dimension-order routing with dimension reversal traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, 2-dimensional reversal traffic, short message, and $q = 1$ (b) 1024 node, 3 virtual channels, 2-dimensional reversal traffic, short message, and $q = 1$, (c) 256 nodes, 3 virtual channels, 3-dimensional reversal traffic, short message, and $q = 1$, (d) 1024 nodes, 3 virtual channels, 3-dimensional reversal traffic, short message, and $q = 1$	97
4.21	Dynamic communication performance of dimension-order routing with bit-reversal traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, short message, and $q = 1$ (b) 512 nodes, 3 virtual channels, short message, and $q = 1$, (c) 1024 nodes, 3 virtual channels, short message, and $q = 1$, (d) 1024 nodes, 3 virtual channels, medium-length message, and $q = 1$, and (e) 1024 nodes, 3 virtual channels, long message, and $q = 1$. . .	98
4.22	Dynamic communication performance of dimension-order routing with complement traffic pattern on various networks: (a) 256 nodes, 3 virtual channels, short message, and $q = 1$, (b) 512 nodes, 3 virtual channels, short message, and $q = 1$, (c) 1024 nodes, 3 virtual channels, short message, and $q = 1$ (d) 1024 nodes, 3 virtual channels, medium-length message, and $q = 1$, and (e) 1024 nodes, 3 virtual channels, long message, and $q = 1$. . .	100
4.23	Dynamic communication performance of dimension-order routing with bit-flip traffic pattern on various networks: (a) 256 nodes, 2 virtual channels, short message, and $q = 1$ and (b) 1024 nodes, 2 virtual channels, short message, and $q = 1$	101
4.24	Dynamic communication performance of large-size HTN by dimension-order routing under various traffic patterns: 3 virtual channels, short message.	102
4.25	Average message latency divided by message length vs. network throughput of HTN: 1024 nodes, 2 VCs, and $q = 1$	103

4.26	Dynamic communication performance of dimension-order routing with different virtual channels and short message on the large-size HTN: (a) hot spot traffic, (b) bit reversal traffic, (c) 2-dimension reversal, (d) 3-dimension reversal, and (e) complement traffic patterns.	104
4.27	Routing messages in a 6×6 mesh from node $(0, i)$ to node $(i, 5)$ (for $0 \leq i \leq 5$); (a) Using dimension order routing, five messages must traverse the channel from $(0, 4)$ to $(0, 5)$, (b) Using adaptive routing, all messages proceed simultaneously.	106
4.28	A 6×6 mesh with a faulty link from node $(3, 2)$ to node $(3, 3)$. (a) With dimension order routing messages from dark nodes to the shaded area cannot be delivered. (b) With adaptive routing, messages can be delivered between all pairs of nodes.	107
4.29	Selection of physical link by link-selection algorithm	108
4.30	Link-selection algorithm for HTN	109
4.31	Selection of virtual channels by channel-selection algorithm	110
4.32	A block diagram of router architecture	114
4.33	Comparison of dynamic communication performance of the HTN between DOR, LS, CS, and LS+CS algorithms with uniform traffic pattern: 1024 nodes, 3 virtual channels, and $q = 1$	119
4.34	Comparison of dynamic communication performance of the HTN between DOR, LS, CS, and LS+CS algorithms with 5% hot-spot traffic pattern: 1024 nodes, 3 virtual channels, short message, and $q = 1$	120
4.35	Comparison of dynamic communication performance of the HTN between DOR, LS, CS, and LS+CS algorithms with bit-reversal traffic pattern: 1024 nodes, 3 virtual channels, 16 flits, and $q = 1$	120
4.36	Comparison of dynamic communication performance of the HTN between DOR, LS, CS, and LS+CS algorithms with bit-flip traffic pattern: 1024 nodes, 3 virtual channels, short message, and $q = 1$	121
4.37	Comparison of dynamic communication performance of the HTN between DOR, LS, CS, and LS+CS algorithms with perfect shuffle traffic pattern: 1024 nodes, 3 virtual channels, short message, and $q = 1$	121
4.38	Dynamic communication performance improvement by LS+CS algorithm over DOR algorithm (a) Maximum throughput enhancement and (b) Message latency reduction.	123
5.1	Hierarchical redundancy of the HTN	127
5.2	Different switch states for reconfiguration: (a) no connect, (b) north-to-south and east-to-west, (c) north-to-west and south-to-east, and (d) north-to-east and south-to-west connects.	127
5.3	Reconfiguration of a plane for the BM in the presence of 4 faulty PEs: Diagonal	128
5.4	Reconfiguration of a plane for the BM in the presence of 4 faulty PEs: Square	128
5.5	Reconfiguration of a plane for the BM in the presence of 4 faulty PEs: Concatenated L-shape and inverse L-shape	129
5.6	Yield for BM and Level-2 network vs. fault density without spare node . .	130
5.7	Yield for BM and Level-2 network vs. fault density with spare node . . .	131
5.8	CONVERGE on a 4×4 2D-mesh	133

5.9	The total number of communication steps of the bitonic merge in different networks	139
5.10	The total number of communication steps of the bitonic merge in different networks	139
5.11	The total number of communication steps of the FFT in different networks	140
5.12	The total number of communication steps for finding the maximum in different networks	140
6.1	A $(4 \times 4 \times 4)$ 3D-torus network (a) unpruned, (b) pruning along the z direction, T_1 , (c) pruning along the $x + y + z$ direction, T_2 , and (d) pruning along the $x + y$ direction, disjoint network.	144
6.2	A (4×4) pruned torus obtained by pruning along the $x + y$ direction. . .	145
6.3	Pruned Hierarchical Torus Network ($m = 4, n = 4$)	146
6.4	An illustration of Level-2 HTN ₃	146
6.5	A comparison of peak number of vertical links of various HTN.	148
6.6	Normalized layout area (1024 PEs, 16 Wafers, and 64 PEs/Wafer)	148
7.1	Basic module of the MH3DT network	151
7.2	Interconnection of a Level-2 MH3DT network	152
7.3	Routing algorithm of the MH3DT network	154
7.4	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: 4096 nodes, 2 VCs, 16 flits	160
7.5	Average transfer time divided by message length versus network throughput of MH3DT network: 4096 nodes, 2 VCs, 16 flits, Buffer Size 2 flits. . .	161
7.6	Dynamic communication performance of dimension order routing with uniform traffic pattern on the MH3DT network: 4096 nodes, various virtual channels,, 16 flits, Buffer Size 2 flits.	162
7.7	Basic module of the TTN	164
7.8	Interconnection of a Level-2 TTN	165
7.9	Routing algorithm of the TTN	166
7.10	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: 4096 nodes, 4 VCs, 16 flits, and $q = 0$	171
7.11	Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: 4096 nodes, 4 VCs, 16 flits, and $q = 1$	171

List of Tables

2.1	A collection of types of interconnection networks used in commercial and experimental parallel computers	12
3.1	Diameter of HTN with Level- L	45
3.2	Comparison of degree and connectivity for various networks	49
3.3	Parameters for layout area in 3D stacked implementation	59
3.4	Comparison of maximum wire length of different networks	62
4.1	The total number of links of various networks with 1024 node	89
4.2	Maximum throughput of the HTN (<i>Flits/Cycle/Node</i>)	102
4.3	Gate counts for router modules	115
4.4	Gate counts for HTN routers	115
4.5	Delays for the router module	115
4.6	Module delay constants for a 0.8 micron CMOS process.	116
4.7	Module delay for a 0.8 micron CMOS process	116
4.8	Performance Improvement using selection algorithm over dimension-order routing	123
5.1	The total number of communication steps on a network for bitonic merge, FFT, and finding the maximum.	138
6.1	Comparison of wiring complexity of various Level-2 HTN	147
7.1	Comparison of static network performance of various network with 4096 node	158
7.2	Comparison of static network performance of various network with 4096 node	169